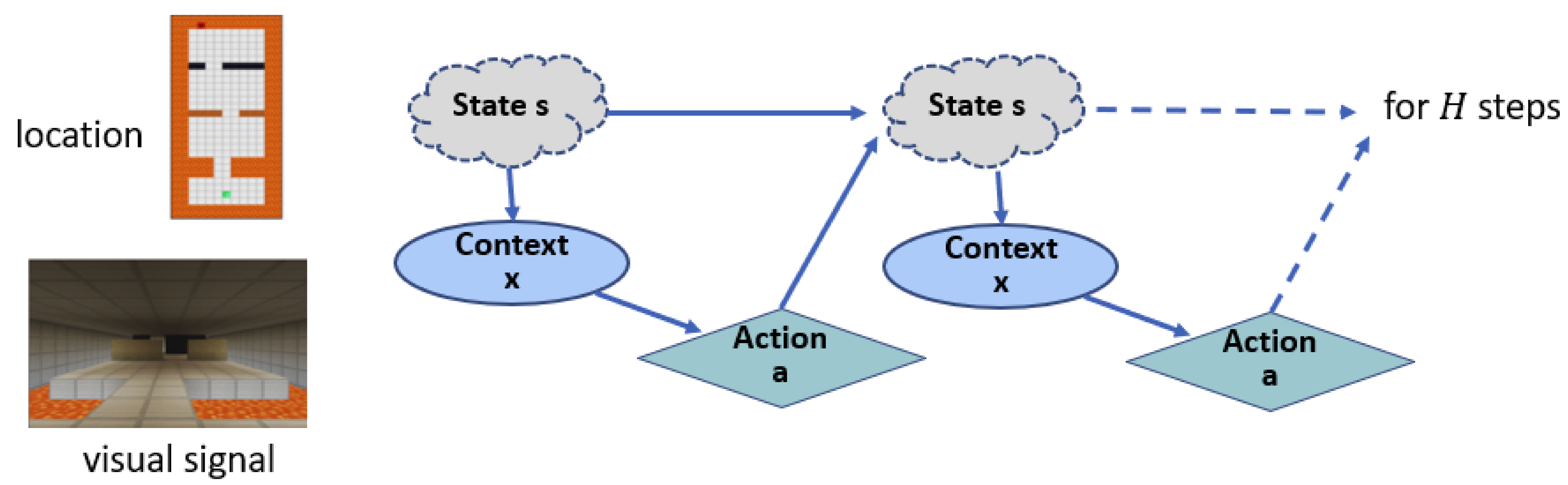
Provably Efficient Exploration for RL with Unsupervised Learning

Fei Feng, Ruosong Wang, Wotao Yin, Simon Du and Lin Yang

fei.feng@math.ucla.edu; ruosongw@andrew.cmu.edu; wotaoyin@math.ucla.edu; ssdu@ias.edu; linyang@ee.ucla.edu

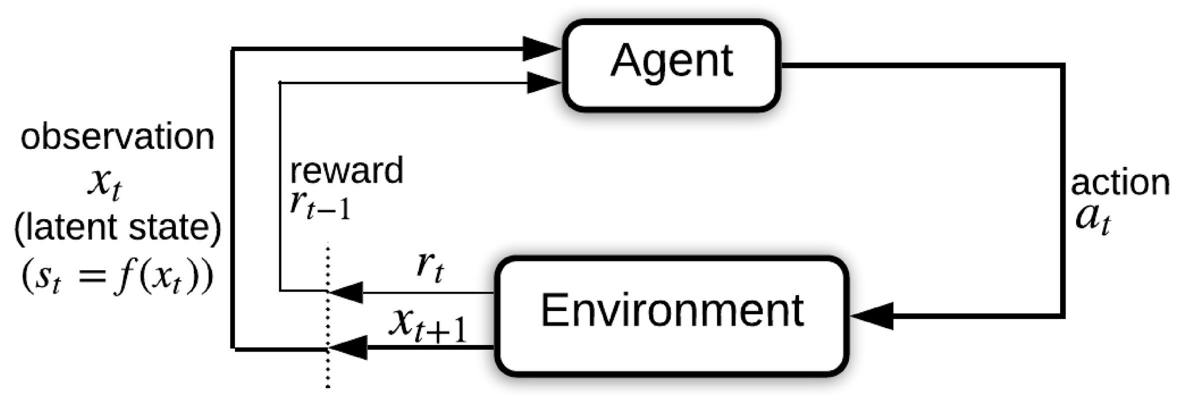
Problem Setting

Reinforcement Learning with Rich Observations



- Rich contexts are generated from a small number of latent states;
- Agent only observes contexts rather than the latent states;
- Studied in e.g., [1, 2]

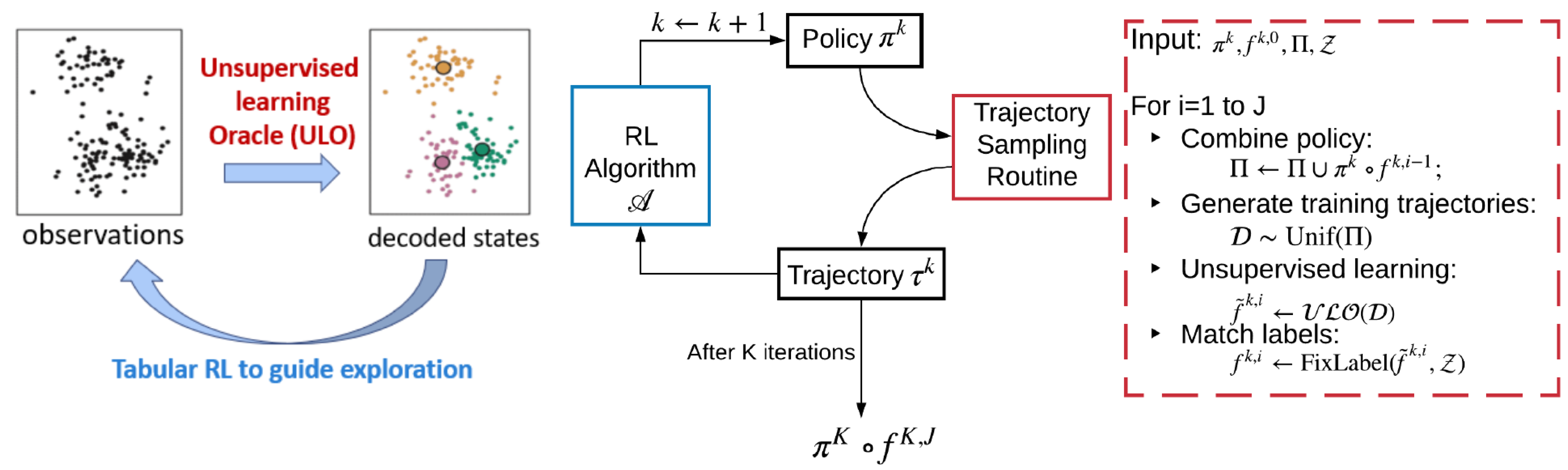
Block Markov Decision Process



$\mathcal{M}' := (\mathcal{S}, \mathcal{A}, \mathcal{X}, P, q, r, f, H)$.

- $\mathcal{X}$ is a huge observation space.
- $\mathcal{S}$ is a small finite latent state space.
- $\mathcal{X} = \cup_{s \in \mathcal{S}} \mathcal{X}_s, \ \mathcal{X}_s \cap \mathcal{X}_{s'} = \emptyset$.
- $x \sim q(\cdot|s), \ f(x) = s$, decoding function.

Our Framework



Unsupervised Learning Oracle $\mathcal{ULO}$

Given n samples $\{x_i\}_{i=1}^n$ generated following $\sum_{s \in \mathcal{S}} q(\cdot|s)\mu(s)$, with probability at least $1-\delta$, $\mathcal{ULO}$ learns a function $\hat{f} : \mathcal{X} \to \mathcal{S}$ such that

$$\mathbb{P}_{s \sim \mu, x \sim q(\cdot|s)}(\hat{f}(x) = s) \geq 1 - g(n, \delta),$$

where $\lim_{n \to \infty} g(n, \delta) = 0$.

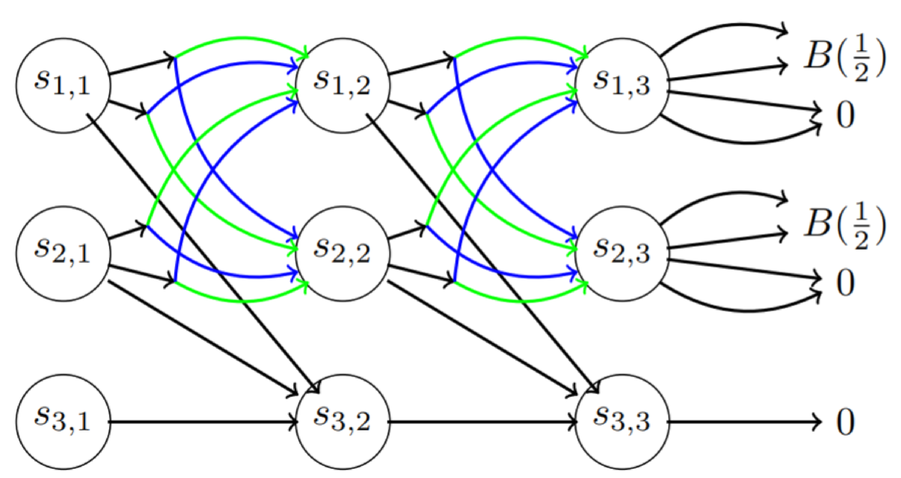
No-regret Tabular RL Algorithm $\mathscr{A}$

For any MDP $\mathcal{M} := (\mathcal{S}, \mathcal{A}, P, r, H)$, $\mathscr{A}$ runs for at most $\mathrm{poly}(|\mathcal{S}|, |\mathcal{A}|, H, 1/\varepsilon, \log(\delta^{-1}))$ episodes to learn an ε-optimal policy with probability at least $1-\delta$.

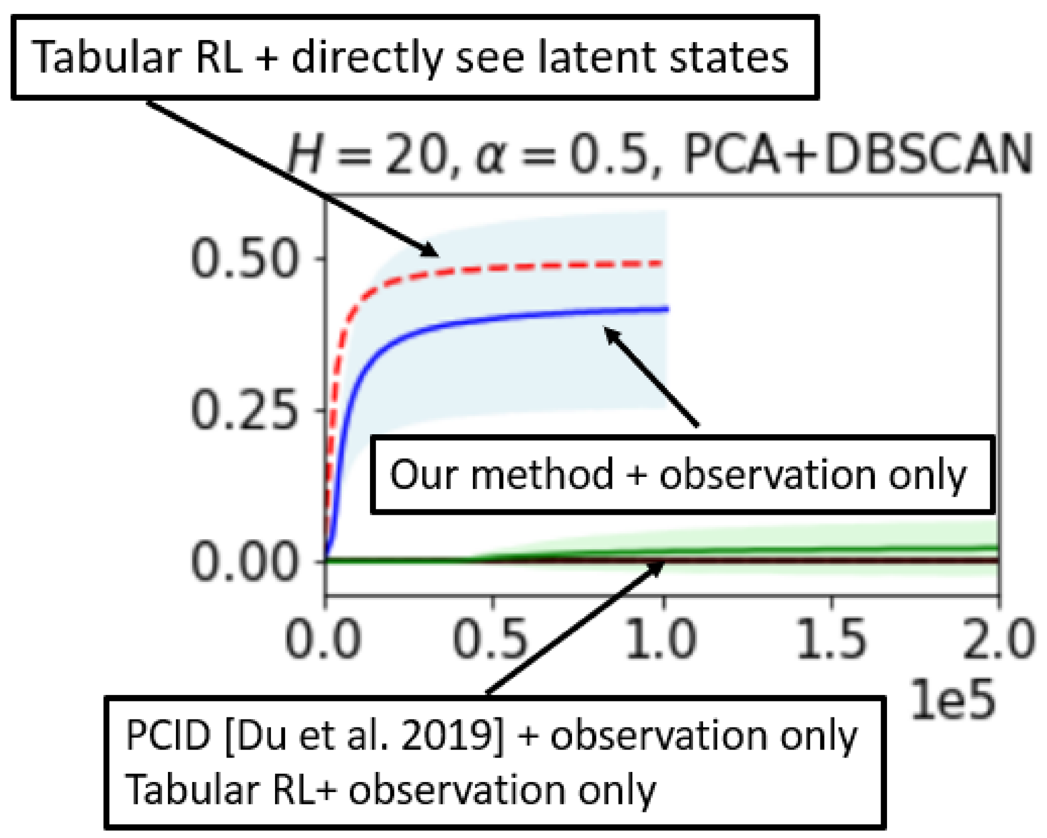
Theoretical guarantee (informal version)

Theorem 1 *Given an efficient $\mathcal{ULO}$ and a no-regret tabular RL algorithm, with at most $\mathrm{poly}\left(|\mathcal{S}|, |\mathcal{A}|, H, 1/\varepsilon, \log(\delta^{-1})\right)$ trajectories, we obtain an ε-optimal policy for the underlying BMDP with probability at least $1-\delta$.*

Numerical Test



- A combination lock environment, hard for random exploration.
- $H+3$-dimensional observation = one-hot encoding of state + H-dimensional noise.



Examples of $\mathcal{ULO}$

- **Gaussian Mixture Models (GMM).** In GMM, states lie in $\mathbb{R}^d$ and observations are states plus some (truncated) zero-mean Gaussian noises.
- **Bernoulli Mixture Models (BMM).** In BMM, observations are points in $\{0,1\}^d$. Every state s is frequency vector $p^s \in [0,1]^d$ such that $q(x|s) = \Pi_{i=1}^d (p_i^s)^{x_i}(1-p_i^s)^{1-x_i}$.
- **Subspace Clustering.** In some cases, each state is a set of vectors and the corresponding observations are points lying in the subspace spanned by the state vectors.

Proper algorithms to serve as $\mathcal{ULO}$ can be found in literature.

Conclusion

We propose a provably efficient framework that turns an unsupervised learning algorithm and a no-regret tabular RL algorithm into an algorithm for RL problems with huge observation spaces.

References

[1] Akshay Krishnamurthy, Alekh Agarwal, and John Langford. Pac reinforcement learning with rich observations. In *Advances in Neural Information Processing Systems*, pages 1840–1848, 2016.

[2] Simon Du, Akshay Krishnamurthy, Nan Jiang, Alekh Agarwal, Miroslav Dudik, and John Langford. Provably efficient rl with rich observations via latent state decoding. In *International Conference on Machine Learning*, pages 1665–1674, 2019.